Draft genome and reference transcriptomic resources for the urticating pine defoliator Thaumetopoea pityocampa

DR. BERNHARD GSCHLOESSL (Orcid ID : 0000-0002-7296-148X)
DR. MATHIEU GAUTIER (Orcid ID : 0000-0001-7257-5880)
DR. CAROLE KERDELHUE (Orcid ID : 0000-0001-7667-902X)

Article type : Resource Article

Draft genome and reference transcriptomic resources for the urticating pine defoliator *Thaumetopoea pityocampa* (Lepidoptera: Notodontidae)

B. Gschloessl[1a], F. Dorkeld[1a], H. Berges[2], G. Beydon[2], O. Bouchez[3], M. Branco[4], A. Bretaudeau[5a,5b], C. Burban[6], E. Dubois[7], P. Gauthier[1b], E. Lhuillier[3], J. Nichols[8], S. Nidelet[7*], S. Rocha[4], L. Sauné[1a], R. Streiff[1a], M. Gautier[1a], C. Kerdelhué[1a]

Full postal addresses

1a- CBGP, INRA, CIRAD, IRD, Montpellier SupAgro, Univ. Montpellier, 755 avenue du Campus Agropolis, CS30016, F-34988 Montferrier-sur-Lez cedex, France

1b- CBGP, IRD, CIRAD, INRA, Montpellier SupAgro, Univ. Montpellier, 755 avenue du Campus Agropolis, CS30016, F-34988 Montferrier-sur-Lez cedex, France

2- INRA-CNRGV, 24 Chemin de Borde Rouge, BP 52627, 31326 Castanet Tolosan Cedex, France

3- INRA, US 1426, GeT-PlaGe, Genotoul, INRA Auzeville, Chemin de Borde Rouge, Auzeville, 31326 Castanet-Tolosan Cedex, France

4- Forest Research Center (CEF), Instituto Superior de Agronomia (ISA), University of Lisbon (ULisboa), Tapada da Ajuda, 1349-017 Lisboa, Portugal

5a- INRA, UMR Institut de Génétique, Environnement et Protection des Plantes (IGEPP), BioInformatics Platform for Agroecosystems Arthropods (BIPAA), Campus Beaulieu, 35042 Rennes, France

5b- INRIA, IRISA, GenOuest Core Facility, Campus de Beaulieu, 35042 Rennes, France

6- BIOGECO, INRA, Univ. Bordeaux, 69 route d'Arcachon, 33610 Cestas, France

7- Plateforme MGX - Montpellier GenomiX, c/o Institut de Génomique Fonctionnelle IGF-sud, UMR 5203 CNRS – U 661 INSERM – Université de Montpellier, 141 rue de la Cardonille, 34094 Montpellier Cedex 05, France

8- Edinburgh Genomics, Ashworth Laboratories, The King's Buildings, The University of Edinburgh, EH9 3FL, Scotland, UK

* Present address: 1a

Keywords (4 – 6): *de novo* assembly, genome, transcriptome, gene prediction, BAC library, Lepidoptera

Corresponding author (name, address, fax, email): Bernhard Gschloessl, CBGP, INRA, CIRAD, IRD, Montpellier SupAgro, Univ. Montpellier, 755 avenue du Campus Agropolis, CS30016, F-34988 Montferrier-sur-Lez cedex, France. Tel.: +33 4 30 63 04 18; fax: +33 4 99 62 33 45. e-mail address: Bernhard.Gschloessl@inra.fr

Running title (45 characters including space): *De novo* genome and transcriptomes of *T. pityocampa*

ABSTRACT

The pine processionary moth *Thaumetopoea pityocampa* (Lepidoptera: Notodontidae) is the main pine defoliator in the Mediterranean region. Its urticating larvae cause severe human and animal health concerns in the invaded areas. This species shows a high phenotypic variability for various traits, such as phenology, fecundity, and tolerance to extreme temperatures. This study presents the construction and analysis of extensive genomic and transcriptomic resources, which are an obligate prerequisite to understand their underlying genetic architecture. Using a well-studied population from Portugal with peculiar phenological characteristics, the karyotype was first determined and a first draft genome of 537 Mb total length was assembled into 68 292 scaffolds (N50=164 kb). From this genome assembly 29 415 coding genes were predicted. To circumvent some limitations for fine scale physical mapping of genomic regions of interest, a 3X coverage BAC library was also developed. In particular, 11 BACs from this library were individually sequenced to assess the assembly quality. Additionally, *de novo* transcriptomic resources were generated from various developmental stages sequenced with HiSeq and MiSeq Illumina technologies. The reads were *de novo* assembled into 62 376 and 63 175 transcripts, respectively. Then, a robust subset of the genome-predicted coding genes, the *de novo* transcriptome assemblies and previously published 454/Sanger data were clustered to obtain a high quality and comprehensive reference transcriptome consisting of 29 701 *bona fide* unigenes. These sequences covered 99% of the CEGMA and 88% of the BUSCO highly conserved eukaryotic genes and 84% of the BUSCO arthropod gene set. Moreover, 90% of these transcripts could be localized on the draft genome. The described information is available via a genome annotation portal (http://bipaa.genouest.org/sp/thaumetopoea_pityocampa/).

INTRODUCTION

The pine processionary moth (hereafter PPM) *Thaumetopoea pityocampa* (Denis & Schiffermüller) is a main pest of Mediterranean pine forests and is widespread in southern Europe and North Africa where it develops at the expense of most *Pinus* and *Cedrus* species (Battisti et al., 2015). The PPM has received increasing attention in the last decades because it is a human and animal health concern, as the highly urticating setae carried by the larvae can cause severe allergic reactions (Battisti, Holm, Fagrell, & Larsson, 2011; Battisti, Larsson, & Roques, 2017; Vega et al., 2014). Moreover, this species is expanding its geographical range due to climate warming, colonizing new, populated areas (Battisti et al., 2005; Robinet et al., 2012; Robinet, Rousselet, & Roques, 2014).

The pine processionary moth is univoltine, and its larvae develop during winter by feeding on needles of several native or introduced conifer species, in forest, agricultural or urbanized areas (Rossi, Garcia, Roques, & Rousselet, 2016). Host preference and larval performance were found to differ between regions (Hodar, Zamora, & Castro, 2002; Zovi, Stastny, Battisti, & Larsson, 2008). The capacity of its larvae to cope with extreme temperatures was also proved experimentally to be a variable adaptive trait (Santos, Paiva, Tavares, Kerdelhué, & Branco, 2011). The timing of sexual reproduction varies as a function of local conditions, adults emerging and mating earlier in colder environments (northern range and high altitudes) compared to warmer places. It has recently been hypothesized that evolution of local phenology as a response to climate change should be taken into account to anticipate future expansion of the PPM (Robinet, Laparie, & Rousselet, 2015). Finally, a population showing an aberrant phenology with adults reproducing in spring and larvae developing during the summer – thus reported as the "summer population" (SP) - was discovered in 1997 in Portugal (Pimentel et al., 2006). This very unique population has been thoroughly studied in the past 10 years. It has been suggested that it probably emerged from a relatively recent

phenological shift, and that gene flow between the SP and the sympatric "classical" winter populations is strongly reduced (Burban et al., 2016; Santos, Burban, et al., 2011).

Although many ecological and phenotypic studies have been conducted so far for the PPM, very little is known about the genomic evolution of its populations. Molecular studies have mostly involved sequencing of mitochondrial or nuclear DNA fragments to unravel phylogeographic patterns at various spatial scales (Kerdelhué et al., 2009; Rousselet et al., 2010), or neutral population genetics approaches using a handful of microsatellite markers to study geographical structure (El Mokhefi et al., 2016), or allochronic differentiation (Santos, Burban, et al., 2011) and introgression patterns (Burban et al., 2016). Recently, transcriptomic resources were developed using Sanger and Roche 454 sequencing technologies (Gschloessl et al., 2014), to provide a first reference of PPM expressed genes with a particular focus on genes potentially involved in phenology. Although incomplete, this reference transcriptome was already useful to compare sets of transcripts between two allochronic populations. New resources were later released to identify associated viral sequences, but most probably corresponded to the sister species *T. wilkinsoni* (Jakubowska et al., 2015). Recent advances in high throughput sequencing technologies now allow to obtain valuable resources for non-model organisms, and to fill the gap between genomic and phenotypic evolution (Mueller, Kuhl, Timmermann, & Kempenaers, 2016). Availability of a reference genome eases large-scale analyses of genetic variation and the development of population genomic approaches to identify genomic regions prone to selection via genome-wide scans (Manel et al., 2016), to disentangle complex demographic histories, or to analyze mosaic of introgression in hybrid zones. In addition to a reference genome, a reference transcriptome further opens the possibility of differential expression studies, and can pave the way to find candidate genes involved in ecologically relevant traits. It represents a useful tool

to annotate genomic sequences, and to interpret polymorphism patterns (Du et al., 2015; Fitak, Mohandesan, Corander, & Burger, 2016). Such resources are urgently needed for various on-going studies concerning the PPM and implying the development of pan-genomic markers to disentangle complex demographic scenarios (Leblois et al., 2018) or to improve the identification (and annotation) of loci subjected to adaptive constraints using genome-wide scans which detect footprints of selection (Gautier et al., pers. comm.). They will also be essential to characterize the genetic architecture of phenology (see for instance Derks et al., 2015), of the urticating system (Berardi et al., 2017) or of traits involved in the adaptation to high temperatures (such as heat shock proteins).

The aim of this study was to provide first insights into the genome and transcriptome of the PPM. Thus, the focus was set on the SP found in Portugal since genetic diversity is lower within this population (Santos, Burban, et al., 2011). This study reports the first karyotype description and *de novo* PPM genome assembly, using BAC sequencing as a quality assessment tool. Furthermore, coding genes predicted on the draft genome and *de novo* assembly of transcriptomic data were combined to obtain a robust and comprehensive reference set of expressed PPM genes that were in turn mapped on the genome.

MATERIAL AND METHODS

Sampling

All the material used in the experiments described below was sampled between 2010 and 2012 from the Mata Nacional de Leiria, Portugal (39°47'N, 8°58'W). *T. pityocampa* SP larvae were directly collected from the nests, pupae were dug out from the soil and males were caught using pheromone baited traps as detailed in Santos, Burban, et al. (2011). Samples from the main developmental stages (adults, eggs, L1, L3 and L5 larvae, freshly buried and metamorphosing pupae) were obtained from laboratory rearing as detailed in

Branco, Paiva, Santos, Burban, and Kerdelhué (2017). Samples used for genomic sequencing were preserved in 95% ethanol, while samples used for BAC construction or RNA extraction were flash-frozen in liquid nitrogen and preserved at -80°C.

Karyotyping

Karyotypes were obtained using fresh eggs washed with a physiological solution (NaCl 0.9%) and following a protocol routinely used for cell lines (Popescu, Hayes, & Dutrillaux, 1998). After centrifugation (400 g for 5 min) and elimination of the supernatant, eggs were placed in RPMI 1640 culture medium with colchicine (0.04 µg/ml final concentration), crushed with a piston pellet and incubated for 3h at room temperature. After centrifugation (400 g for 5 min), supernatant was eliminated and the pellet resuspended in hypotonic solution (KCl 0.075 M) and incubated for 10 min at room temperature. Mitotic chromosomes were then fixed twice with a methanol:acetic acid solution (3:1), spread on slides and counted under a fluorescent microscope following DAPI staining.

BAC library construction

High molecular weight DNA was prepared from 130 L1 larvae hatched in the laboratory from egg masses collected in the field. The protocol described in Peterson, Tomkins, Frisch, Wing, and Paterson (2000) and Gonthier et al. (2010) was applied with the following modifications: (1) Sucrose based Extraction Buffer (SEB) was 0.01 M Tris, 0.1 M KCl, 0.01 M EDTA pH 9.4, 500 mM sucrose, 4 mM spermidine, 1 mM spermine tetrahydrochloride, 0.1% w/v ascorbic acid, 0.25% w/v PVP 40 000, and 0.13% w/v sodium diethyldithiocarbamate, (2) lysis buffer was 1% w/v sodium lauryl sarcosine, 0.1 mg/ml proteinase K, 0.13% w/v sodium diethyldithiocarbamate, 6 mM EGTA and 200 mM L-Lysine dissolved in 0.5 M EDTA pH

9.1, (3) after lysis of the nuclei, agarose plugs were pre-washed 1h in 0.5 M EDTA pH 9.1 at 50°C, 1h in 0.05 M EDTA pH 8 at 4°C, and then stored at 4°C. Partial digestion of high molecular weight genomic DNA with *Hind*III (Sigma-Aldrich, St-Louis, Missouri), elution and ligation to pIndigoBAC-5 *Hind*III-Cloning Ready vector (Epicentre Biotechnologies, Madison, Wisconsin) were performed according to Chalhoub, Belcram, and Caboche (2004). The BAC library was deposited at the Centre National de Ressources Génomiques Végétales (Toulouse, France).

Sequencing and assembly of 11 BACs for assembly quality assessment

DNA was isolated from 11 randomly-chosen BACs using the Nucleobond Xtra Midi Plus kit (Macherey Nagel) following the manufacturer instructions using 100 mL of LB media with 12.5 µg/mL Chloramphenicol selective marker. Then, 15 µg of DNA were obtained for each BAC clone (ca. 150 ng/µL), of which 150 ng were digested with the fast NotI enzyme (Fermentas) and incubated 40 min at 37°C. After incubation, the enzymatic digestion was transferred on a 0.8% agarose gel (TBE 0.25X) for pulse field electrophoresis performed with a Chef Mapper XA CHILLER SYSTEM 220V (Bio-Rad) under the following conditions: voltage = 6 v/cm, included angle = 120°, initial switch time = 5 sec, final switch time = 15 sec, run time = 16 hours, ramping = linear. Each insert size was estimated using the software GeneTools (Syngene).

Libraries were constructed with the TruSeq Nano DNA sample preparation (low throughput protocol) kit from Illumina. 250 ng of BACs were fragmented through sonication on a Covaris S220 to produce fragments of ca. 900 bp. The DNA fragments went through an end repair process, and were purified and size selected on magnetic beads. Finally, a single 'A' base was added. The adapter and a molecular index were subsequently ligated. The DNA fragments were selectively enriched with 8 PCR cycles. The final DNA libraries were

validated with a DNA 1000 Labchip on a Bioanalyzer (Agilent) and quantified with a KAPA qPCR kit. After normalization, libraries were pooled in equimolar amounts and sequencing was performed on an Illumina MiSeq using the paired-end (PE) protocol and the Reagent Kit v3 (2x300 bp).

The raw PE reads were cleaned with the Trimmomatic software (v0.33, Bolger, Lohse, & Usadel, 2014) to remove the Illumina adapter as well as the vector sequences, using the following parameters: ILLUMINACLIP:$adapterfile:2:28:10 ILLUMINACLIP:$vectorfile:2:28:10 HEADCROP:15 LEADING:28 TRAILING:28 SLIDINGWINDOW:5:30 MINLEN:30. Each BAC assembly (Fig. 1A) was established using Velvet (v1.2.10, Zerbino & Birney, 2008) on both the cleaned PE reads (applying an inner distance between read1 and read2 of 300 bp) and the Single-End (SE) reads reported as 'unpaired' by Trimmomatic. Several assemblies were created using k-mer sizes from 31 to 261 with steps of 10. The chosen assembly corresponded to the highest value of the product N50 [bp] x assembly size [bp] weighted by the number of contigs. The best assembly was aligned against NCBI NT (v07/2015, NCBI Resource Coordinators, 2016) to search for diverse contaminations (*e.g.* bacteria, nematodes) using *blastn* BLAST (Altschul, Gish, Miller, Myers, & Lipman, 1990) web-analysis via http://www.blast.ncbi.nlm.nih.gov/Blast.cgi. Contaminated contigs were identified and removed by analyzing manually all alignment results having e-values between e-14 and 0, high bit scores and corresponding to non-insect species sequences. Contaminated contigs matched mainly to *Actinobacteria*, *Proteobacteria* (both with e-values of 0) and flatworms (*Platyhelminthes*, e-value of e-14). Other contigs were identified as being artificial sequences (*i.e.* cloning vectors), uncultured bacteria (*i.e.* adhering to rumen of cows) and a sequence matching best to *Physeter catodon* (sperm whale). The remaining contigs for each BAC were then scaffolded with the program SSPACE (v3.0, Boetzer, Henkel, Jansen, Butler, &

Pirovano, 2011). Two independent scaffolding strategies were applied. The first scaffolding analysis included only the PE sequences (applying standard parameters: -k 5 –a 0.7 –x 0 –p 1) while the second approach also took into account the Trimmomatic SE reads (same parameters). This allowed SSPACE to extend contigs when possible. Scaffolds shorter than 1500 bp were removed from each assembly. Then, as above, the best of these two scaffoldings was chosen by calculating the (N50 [bp] x assembly size [bp] / contig number) criterion.

The quality of the 11 assembled BACs was assessed by screening each BAC for the presence of duplicated regions. Thus, each assembly was aligned against itself by global alignment with LASTZ (v1.03.02, Harris, 2007) applying the following parameters: --notransition --step=20 --gfextend --gapped --chain --matchcount=800 --identity=92 --format=general:score,name1,zstart1,end1,strand1,size1,name2,zstart2+, end2+,strand2,size2.

Read coverage was determined by mapping the Trimmomatic output reads against each BAC scaffold using Bowtie (v2.2.4, Langmead & Salzberg, 2012). In addition to default parameters and --reorder, only read pairs with mapping distances consistent with the specific library insert size were retained, *i.e.* applying the options --no-discordant and --maxins 1100. Subsequently, the aligned read percentage and the read per bp coverage were calculated for each BAC assembly using SAMtools (v1.2, H. Li et al., 2009), BAMtools (v2.3.0, Barnett, Garrison, Quinlan, Stromberg, & Marth, 2011) and BEDtools (v2.2.23, Quinlan & Hall, 2010).

Construction and quality assessment of the *Tpit*-SP v1 genome assembly

Genome library construction and sequencing

Whole genomic DNA was isolated following a standard CTAB/phenol chloroform protocol from the head and thorax of 21 SP males caught by pheromone trapping. Six libraries, hereafter referred to as PE300i, PE600i, SE454, LJD3i, LJD8i and LJD20i, were further constructed. The PE300i and PE600i consisted of Illumina PE libraries with short-insert sizes of 300 and 600 bp respectively, and were constructed based on the genomic DNA from a single male. These two libraries were sequenced on a single lane of an Illumina HiSeq2000 in the Edinburgh Genomics sequencing facility (Edinburgh, Scotland). The four other libraries were constructed and sequenced by Eurofins MWG Operon (Ebersberg, Germany). In short, the SE454 library consisted in a whole genome SE library, built from a single male and sequenced on 3 runs of a Roche 454 GS FLX+. The LJD3i, LJD8i and LJD20i libraries were Long Jumping Distance (LJD) PE libraries with insert sizes of 3000, 8000 and 20 000 bp, respectively. LJDs were developed by Eurofins MWG Operon as an alternative to mate pair sequencing; they have larger insert sizes than mate pair libraries and are sequenced using the paired-end read protocol (https://www.eurofinsgenomics.eu/en/eurofins-genomics/product-faqs/next-generation-sequencing.aspx). Each library was constructed using a pool of 4 to 8 males. Pooling of DNAs was necessary to obtain enough DNA at the unwanted expense of increasing genetic diversity in the sequenced sample. These three libraries were sequenced on a single Illumina HiSeq2000 lane.

De novo *genome assembly and characteristics*

Assembly of all the resulting raw data was carried out by Eurofins MWG Operon. Briefly, the reads of all libraries were cleaned for quality and adapters were removed (Fig. 1B) with

the Trimmomatic software (v0.22, window size of 20 bp, minimum PHRED quality of 20, Bolger et al., 2014). PE reads were filtered if they were shorter than 70 bp, LJD reads if they were shorter than 30 bp and SE454 reads if they were shorter than 100 bp. Furthermore, reads mapping to a bacterial genome in the NCBI Genome database were considered as contaminations and removed. Please note that this procedure might have resulted in the potential removal of endosymbiont genes integrated in the nuclear genome. Subsequently, Eurofins MWG Operon applied their own assembly pipeline (CONVEY, http://www.conveycomputer.com) which ran multiple Velvet assemblies with all available Illumina and 454 data using a broad range of k-mer sizes and varying parameters. The best assembly (k-mer size 61) was selected on the basis of N50 size and maximal scaffold length. Scaffolds with a length of at least 1000 bp were retained to constitute the final genome assembly of *T. pityocampa*, named *Tpit*-SP v1 genome. The mitogenome was identified and isolated from the nuclear genome assembly. The AT-rich sequence of the mitochondrion was detected in a single genome scaffold of 15 717 bp, confirmed by *blastn* against the NCBI Nucleotide database and further manually edited. Furthermore, the circular structure of the mitogenome was verified.

Classical scaffold length statistics (N50, N90, mean, *etc.*) and the GC content of the nuclear genome were calculated on the assembled scaffolds. In order to evaluate genome coverage, the PE reads of the three genomic libraries were mapped on the contigs and scaffolds with Bowtie (as described above for BAC coverage), applying --maxins 500 bp for PE300i and 800 bp for PE600i, respectively. The SE454 reads were mapped in single-end mode (Bowtie parameter -U). In addition, to estimate *in silico* the expected full genome size of the *Tpit*-SP v1 genome, the frequencies of kmers of 61 bp length were counted within all Illumina paired-end raw reads with the software Jellyfish (v1.1.11, Marcais & Kingsford, 2011) and the

results were subsequently analyzed with the web-software GenomeScope (parameters: kmer=61, max cov=80, read length=100bp; Vurture et al., 2017).

Furthermore, the same LASTZ approach (see above) was applied as for the BAC analyses to identify duplicated regions and haplotypes possibly assembled in different scaffolds. Repeats in the genome assembly were identified with RepeatModeler (v1.0.8, Smit & Hubley, 2008-2015) (parameter: -engine ncbi) and RepeatMasker (v4.0.6, Smit, Hubley, & Green, 2013-2015) (parameters: -s xsmall -gff –norna –engine ncbi –poly –gff). The results of RepeatMasker were summarized by the tool "One code to find them all" (v03/2016, Bailly-Bechet, Haudry, & Lerat, 2014) applying default parameters.

Quality assessment of the Tpit-*SP v1 genome assembly*

The completeness of the assembly was assessed by detecting well-conserved eukaryotic core genes, using (i) the CEGMA prediction pipeline (v2.5, default parameters, Parra, Bradnam, & Korf, 2007) which searches for 248 orthologous groups of proteins (KOGs, Tatusov et al., 2003), and (ii) the recently developed BUSCO program (v1.2, Simao, Waterhouse, Ioannidis, Kriventseva, & Zdobnov, 2015), which was separately run to search for 429 conserved eukaryotic genes (parameters -l eukaryota -m genome --long) and 2675 conserved arthropod genes (parameters -l arthropoda -m genome --long).

CEGMA and BUSCO protein sequences which were not identified in the *T. pityocampa* genome were extracted from the CEGMA (*kogs.fa*) and BUSCO (corresponding *ancestral* FASTA files for 'eukaryota' and 'arthropod' analyses) KOG sequences and subsequently aligned with *tblastn* (NCBI-BLAST+ v2.2.29; parameters: -evalue 1e-5 –outfmt 7 –num_alignments 20) to the genome assembly. The corresponding KOGs were allowed to be

split over several *T. pityocampa* scaffolds, using an in-house Perl script. A KOG was counted as being present if at least 40% of the KOG sequence was aligned on two scaffold ends.

To further assess the quality of the genome assembly, LASTZ was run to align the genome scaffolds (same parameters as mentioned above) as well as the genome contigs (parameter --matchcount=400 in order to take into account the smaller size of contigs) against the 11 BAC assemblies. In addition the cleaned genomic shotgun reads from the libraries PE300i and PE600i were mapped with Bowtie on the BAC sequences in order to calculate the read per base coverage, applying the same protocol and parameters as described for the genome coverage analyses. The alignments of the *Tpit*-SP v1 scaffolds on the BAC assemblies were visualized separately for each BAC using the CIRCOS toolkit (v0.68, Krzywinski et al., 2009).

Construction and quality assessment of *Tpit*-SP transcriptomic resources

Molecular procedures

RNA was extracted from samples corresponding to various developmental stages (eggs, L1, L3 and L5 larvae, freshly buried and metamorphosing pupae, adults) using a Trizol extraction procedure. RNA quality and integrity was evaluated through migration on an agarose gel and nanodrop technology. Whenever the 260/280 OD ratio was below 1.7, samples were purified using the Qiagen RNeasy plant mini kit. RNA concentrations were estimated using the Qubit procedure (Quant-it RNA assay kit). RNAs from 2 to 5 individuals were pooled for each development stage in equimolar proportions before library preparation.

Libraries were constructed using the TruSeq stranded mRNA sample prep kit (Illumina) according to the manufacturer instructions, both with a standard insert size of 450 bp and with a long insert size of 850 bp obtained by modifying the fragmentation time and beads

quantity. Briefly, poly-A RNAs were purified using oligo-d(T) magnetic beads. The poly-A+ RNAs were fragmented and reverse transcribed using random hexamers, Super Script II (Life Technologies) and Actinomycin D. During the second strand generation step, dTTP was substituted by dUTP. This prevented the second strand to be used as a matrix during the final PCR amplification. Double-stranded cDNAs were adenylated at their 3' ends before ligation was performed using Illumina's indexed adapters. Ligated cDNAs were amplified through 15 PCR cycles and PCR products were purified using AMPure XP Beads (Beckman Coulter Genomics).

Libraries were validated using a DNA1000 chip on an Agilent Bioanalyzer and quantified using the KAPA Library quantification kit (Clinisciences). The libraries obtained using 450 bp inserts were pooled in equimolar amounts and sequenced in two lanes of an Illumina HiSeq2000 using the PE 2x100 bp protocol. Then, all libraries (both standard and long insert sizes) were pooled and sequenced using the MiSeq PE 2x300 bp protocol.

Building a predicted gene set from the Tpit-*SP v1 genome*

The AUGUSTUS program (Stanke & Waack, 2003) was used to identify potential coding regions in the genome. This program relies on Markov models to represent and predict the gene structure in a specific genome. Due to the lack of a Lepidoptera-specific coding region model, a *de novo* gene model for *T. pityocampa* was created. Thus, WebAUGUSTUS (Stanke & Morgenstern, 2005) was run in the 'Training' mode, providing the assembled genome scaffolds and 43 486 Lepidoptera protein sequences obtained from the UniRef50 database (redundancy removed at 50% identity at maximum, Suzek, Wang, Huang, McGarvey, & Wu, 2015). The generated training model was retained for the gene prediction analysis.

To optimize the identification of coding regions, the Illumina HiSeq RNAseq reads were included as extrinsic data (*i.e.* evidence from external sources) into the prediction process.

These reads were mapped with TopHat (v2.0.12, Kim et al., 2013), applying parameters -r 100 and --no-discordant. Following the AUGUSTUS standard recommendations for eukaryotic genomes (Stanke, 2009), the genome-mapped RNAseq read clusters were identified as potential exons and the inter-exonic regions as potential introns. In order to obtain the most specific coding-gene structure annotation, two consecutive AUGUSTUS (v3.0.2) predictions (Fig. 1B) were applied with the following parameters as suggested in Stanke (2014): --species=$NEW_PPM_MODEL --extrinsicCfgFile=extrinsic.M.RM.E.W.cfg --alternatives-from-evidence=true --allow_hinted_splicesites=atac --protein=on --exonnames=on --codingseq=on $GENOME_FASTA. The first prediction (Aug2.1, including the parameters --hintsfile=$ALL_HINTS --introns=on --genemodel=complete) identified only complete coding genes and detected introns and coding sequences (CDS). This gene prediction represented the most exhaustive set of coding genes for the present *Tpit*-SP v1 draft genome. A second AUGUSTUS prediction (Aug2.2) was defined using more robust and conservative criteria before being included to build the *T. pityocampa* reference transcriptome (see below). In detail, the intron genome coordinates were extracted from the Aug2.1 gene structure file (GFF3 format) to retrieve the FASTA sequences of the exon-exon junctions by taking 100 bp of each flanking exon. The HiSeq RNAseq reads were then mapped with Bowtie on these junctions as SE reads ('-U' option) to optimize the exon border reconstruction. Uniquely mapped exon-exon reads retrieved by SAMtools were combined using BAMtools with the uniquely exon-mapped reads previously aligned by TopHat (excluding exon-exon mapping reads). The intron structure coordinates file was updated (*bam2hints* tool provided with AUGUSTUS) and applied as input for the Aug2.2 prediction (including the parameter --hintsfile=$INTRON_HINTS).

Finally, the CDS sequences of the Aug2.1 and Aug2.2 predictions were extracted by applying the AUGUSTUS tool *getAnnotFasta.pl*.

De novo *assembly of HiSeq and MiSeq transcriptomes*

The HiSeq and MiSeq reads were independently analyzed by Trimmomatic for quality filtering and adapter trimming with the following parameters: ILLUMINACLIP: $adapterfile:2:40:15 HEADCROP:12 SLIDINGWINDOW:4:15 MINLEN:30). Subsequently, prinseq-lite (v0.20.2, Schmieder & Edwards, 2011) (-trim_tail_left 5, -trim_tail_right 5, -min_len 30, -out_format 3) was applied to remove polyA and polyT ends longer than 5 bp.

HiSeq and MiSeq data were assembled separately (Fig. 1C) in order to avoid creating chimeric transcripts or artefacts related to the different sequencing technologies. In each case, overlapping reads were combined with the program FLASH (v1.2.11, Magoc & Salzberg, 2011). Merged as well as unmerged FLASH output reads were retained for the assembly process. To decrease computation time and to facilitate the assembly process for the HiSeq reads, redundancy was removed using the BBNorm (v35.21, Bushnel, 2014) normalization tool. Subsequently, Velvet (v1.2.08, default parameters) was run followed by Oases (v0.2.08, default parameters, Schulz, Zerbino, Vingron, & Birney, 2012). Odd k-mer values ranging from 27 to 81 were applied for HiSeq, and from 51 to 101 for MiSeq data, using a step size of 4. In both cases Velvet tool *velvetg* (parameters -amos_file yes -read_trkg yes) was run to determine the insert length to be used in Oases, leading to 220 bp for HiSeq data and 210 and 460 bp for the two MiSeq libraries, respectively. Interestingly, for both MiSeq libraries the *in silico*-estimated insert sizes were shorter than expected from the wet laboratory procedures. This could be due to the relatively wide and flat curve of DNA fragment insert sizes obtained from the Agilent Bioanalyzer quality check (data not shown), which most probably resulted in the over-representation of smaller fragments in the libraries. A comparative analysis (not shown) between Velvet/Oases assemblies established with the wet lab insert sizes further showed that the *in silico*-based insert sizes returned better transcript assemblies. The best

transcriptome for each data set was chosen based on the highest ratio of identified CEGMA core genes, including partial ones. Then, transcripts were filtered on their abundance with the RSEM (v1.2.8, B. Li & Dewey, 2011) program which was run by the tool *align_and_estimate_abundance.pl* (--est_method RSEM) embedded in the Trinity package. Each read was aligned with Bowtie (v2.2.4, default parameters) on the reconstructed transcripts of the corresponding transcriptome. Guided by these alignments, RSEM estimated the abundance of each transcript. For this purpose, the metric, `fragments per kilobase transcript length per million fragments mapped` (FPKM) was chosen to further exclude transcripts with low coverage (*i.e*. FPKM ≤ 2). The read per base coverage and the remapping percentage were calculated for each transcriptome as described in the BAC coverage calculation section.

Establishing a reference transcriptome from the de novo *transcriptomes and the predicted gene set*

To generate a consistent reference set of protein-coding transcripts, first the coding regions of the assembled transcripts were identified. FrameDP (v1.2.2, default parameters, Gouzy, Carrère, & Schiex, 2009) was applied on each of the MiSeq and HiSeq transcripts, on the previously published 454/Sanger transcriptome resource obtained for the SP (Gschloessl et al., 2014) and on the CDS of the Aug2.2 subset of coding genes. For each sequence including a coding region identified by FrameDP, only the longest peptide was retained to limit redundancy in the final data set. This set of protein sequences was then clustered by CD-HIT (CD-HIT package, v4.5.4, Fu, Niu, Zhu, Wu, & Li, 2012) using an identity of 90% (parameters -c 0.90 -l 20), and the results were visualized by a Venn diagram using the R package VennDiagramm (v1.6.17, H. Chen & Boutros, 2011). Again, only the longest protein sequence of each CD-Hit cluster was retained to build the final set of coded reference

proteins. Then, the reference *Tpi*t-SP transcriptome was obtained by retrieving the coding nucleotide sequence corresponding to each protein.

Quality assessment of the two de novo *and the reference transcriptomes*

The completeness of the HiSeq and MiSeq transcriptome assemblies as well as the reference transcriptome was assessed by running CEGMA and BUSCO (using the eukaryote and the arthropod sets) analyses as described above except that the parameter --long was omitted and -m trans was applied as the analysis was done on a transcript set. The ortholog hit ratio (OHR, O'Neil et al., 2010) was also calculated, which estimates the completeness of an assembled transcript based on its *blastx* alignment length compared to the best match within a protein reference set. The HiSeq, MiSeq and reference transcripts were aligned with *blastx* (outfmt 5 -evalue 1e-5 -gapopen 11 -gapextend 1 -word_size 3 -matrix BLOSUM62) against a reference set corresponding to 471 938 protein sequences extracted from the lepidopteran genome database Lepbase (v4, Challis, Kumar, Dasmahapatra, Jiggins, & Blaxter, 2016). The OHR values were then calculated on the best hit sequence as described in Gschloessl et al. (2014).

Reference transcript localization in the *Tpit*-SP v1 genome and functional annotation

Mapping the reference transcripts onto the genome assembly

The transcripts from the three established transcriptomes (HiSeq, MiSeq and reference) and the published 454/Sanger resource obtained from the same population were aligned to the *de novo* genome assembly using BLAT (v35, default parameters, Kent, 2002). As proposed by Stanke (2009), only those transcripts which were aligned to the genome with at least 80% of the transcript length at a 92% sequence identity were kept. Furthermore, transcripts

potentially split over several scaffolds were searched for, applying the same approach as for the quality assessment of genome assembly. The reference transcripts were aligned with *blastn* (parameters: -soft_masking true –evalue 1e-5 –outfmt 7 –show_gis –perc_identity 70) against the *Tpit*-SP v1 assembly. Then, these transcripts were declared as being split on two scaffolds if at least 40% of the transcript was aligned on two scaffold termini.

Functional annotation

The proteins corresponding to the reference transcripts were predicted using FrameDP (default parameters). If multiple proteins were identified for a specific transcript, only the longest peptide sequence was kept. Then, the retained PPM proteins were aligned against the NCBI NR database (v5 june 2017) with *blastp* (v2.5.0, minimum e-value=1e-8, maximum target sequences=20). An InterProScan (v5.13-52.0, Zdobnov & Apweiler, 2001) was conducted on the same dataset. Subsequently, Blast2GO (v2.5, database vOct2016, Conesa et al., 2005) assigned Gene Ontology (GO) terms to proteins, taking into account the *blastp* and InterProScan results. Finally, the OrthoMCL package (L. Li, Stoeckert, & Roos, 2003) (blastp 2.5.0 parameters: -evalue 1e-5, orthomcl v2.0.9 parameters: percentMatchCutoff=50 evalueExponentCutoff=-5, mcl v14-137 with parameters: --abc -I 1.5) was used to identify potential orthologs between the FrameDP-predicted proteins corresponding to the reference transcript set and the protein sets of *Drosophila melanogaster* (version dmel-all-translation-r6.03), *Danaus plexippus* (version Danaus_plexippus.DanPle_1.0.25), *Bombyx mori* (version http://sgp.dna.affrc.go.jp/ComprehensiveGeneSet) and the noctuoid *Spodoptera frugiperda*. For this latter species, the transcript set published by Legeai et al. (2014) was retrieved which is available in LepidoDB (http://bipaa.genouest.org/data/public/lepidodb/TR2012b.fa), and the corresponding predicted protein set was generated with FrameDP (default parameters).

Finally, the longest protein sequence was kept for each transcript with a predicted CDS and these 22 253 *S. frugiperda* proteins were included into the OrthoMCL ortholog analysis.

Public release of the *T. pityocampa* SP resources in LepidoDB database

The genomic and transcriptomic resources developed in the present study were integrated into LepidoDB, a public database hosting genomic resources for several lepidopteran species, which can be accessed via http://bipaa.genouest.org/sp/thaumetopoea_pityocampa/. In detail, all predicted genes (Aug2.1 and Aug2.2 sets) and the reference transcripts along with their corresponding functional annotations were loaded into a Chado database (v1.31, Mungall & Emmert, 2007). The available data for the reference transcript set were gathered in specific web pages using an in-house J2EE application. For visualization purposes, a JBrowse (v1.12.1, Skinner, Uzilov, Stein, Mungall, & Holmes, 2009) genome browser was set up. To facilitate analyses of the PPM resource data, a search engine as well as a BLAST and a Galaxy server (Blankenberg et al., 2010) can be accessed from the web page.

RESULTS AND DISCUSSION

Karyotyping

A diploid number of 98 chromosomes was observed in 6 of the 8 detected metaphases (Fig. S1 Supporting Information), and slightly smaller diploid numbers in the 2 others. Because of the difficulty in spreading the small chromosomes of *T. pityocampa*, it was considered that the actual number was likely 2n=98, while the observations of lower counts possibly reflected incomplete spreading. The small number of observed metaphases may be due to the use of a protocol primarily developed for cell line cultures. Yet, the results were sufficient to

provide a first estimate of chromosome number for the studied species, and to compare them with other Lepidoptera. Indeed, in Lepidoptera, chromosome numbers range from n=5 to n=223 haploid chromosomes, with a majority of taxa showing a constant number of n=31 (*i.e.* chromosomal conservatism) (Ahola et al., 2014; Lukhtanov, 2014), which is supposed to correspond to the ancestral Lepidoptera karyotype. As in other lepidopteran species, the relatively large number of small chromosomes (n=49) in the pine processionary moth is probably due to chromosomal rearrangements and high levels of repetitive elements (Ahola et al., 2014; Lukhtanov, 2014).

Characteristics and quality of the *Tpit*-SP v1 genome

The *T. pityocampa* SP genome was assembled into 675 934 contigs representing 507 Mb (Table 1). N50 and N90 contig lengths of the assembly were 1424 and 320 bp, respectively. These contigs were further assembled into 68 292 scaffolds. The *Tpit*-SP v1 scaffold lengths ranged from 1 kb to 2.1 Mb and the corresponding N50 and N90 scaffold lengths were 164 kb and 1951 bp, respectively. Overall statistics of the genome assembly are listed in Table 1. More details on all libraries shotgun and LJD libraries can be found in Table S1 (Supporting Information). Relatively high scaffold numbers have also been reported most recently for other *de novo* lepidopteran genome assemblies which resulted in 142 to 80 479 scaffolds, and N50 lengths ranging from 5.2 kb to 10.7 Mb (Table 2). The *Tpit*-SP v1 genome size was 537 Mb, which is close to the maximum documented for Notodontidae (estimated range 323 – 587 Mb, Gregory et al., 2007). It is larger than the genome of the Fall armyworm (corn strain: 438 Mb, rice strain: 371 Mb), *S. frugiperda* (Gouin et al., 2017), which is a main pest of rice and corn and the closest Noctuoidea species for which a genome was previously available. The size of the *Tpit*-SP v1 assembly lies in between the ones published for the silkworm *Bombyx mori* (482 Mb) and the winter moth *Operophtera brumata* (638 Mb). Yet, using

GenomeScope, the expected full genome size was estimated to range between 432 and 452 Mb, which suggests that the actual *T. pityocampa* genome size is probably shorter than the *Tpit*-SP V1 assembly size. The mean coverage in the *Tpit*-SP v1 genome assembly was 84 reads/bp (see Table S2 for average coverage per scaffold). All positions of the draft genome assembly (excluding positions with Ns) had a read per base coverage of at least 1. Moreover, 98% of the *Tpit*-SP v1 assembly had a coverage of at least 10, 85% of at least 30 and 39% of at least 100 reads per base pair. These findings confirmed the high quality assembly of the genome contigs. The *Tpit*-SP assembly also had a GC content of 37% which is very close to the other Noctuoidea species, *S. frugiperda* (36% for both strains, Gouin et al., 2017). However, the proportion of repeated elements of 45% (Table 3) was higher than those found for *S. frugiperda* (29% for both strains, Gouin et al., 2017) and closer to the values found for *B. mori* (44%) and *O. brumata* (54%) (Derks et al., 2015; International Silkworm Genome, 2008). Furthermore, 605 duplicated regions (average length: 1348 bp, range 809 - 8509 bp) were identified in 540 scaffolds. Lepidoptera with a larger number of chromosomes are supposed to have shorter chromosomes and a higher level of repeated elements than their counterparts with lower chromosome numbers (Ahola et al., 2014), which is consistent with the *T. pityocampa* genome characteristics of the present study.

Furthermore, the quality of the *Tpit*-SP v1 genome assembly was evaluated by localizing the scaffolds on the reconstructed BACs. In short, the BAC library was composed of 19 968 clones with a mean insert size of 75 kb and represented 3 genome equivalents. Details on the paired-end libraries can be found in Table S3. The randomly chosen 11 sequenced BACs were assembled into 1 to 9 scaffolds (on average 3 scaffolds). Assembly sizes were consistent with the size estimates obtained during the library construction process. The read coverage of the BAC assemblies varied from 921 to 2148 reads/bp (Table 4). When aligning the *T. pityocampa* genome scaffolds against the assembled BACs, all BACs could be recovered on

average at 56% of their length (min. 14%, max. 83%). Between 5 and 32 genome scaffolds (on average 20 scaffolds) were aligned to each BAC sequence (Fig. 2: 4 chosen BACs, Fig. S2: all 11 BACs). On the other hand, the *Tpit*-Sp v1 contigs covered on average 74% of the BAC sequences. In addition, 71 to 93% of the BAC lengths were covered by at least 10 reads per base pair when genomic cleaned reads were mapped (Table 4). Once more, these results illustrated that the *Tpit*-SP v1 genome assembly was of good sequence quality but remained fragmented. At last, the number of conserved eukaryotic and arthropod genes being recovered in the assembly was determined. Among the 248 conserved eukaryotic genes searched by CEGMA, 145 (59%) were identified in the genome assembly, including 87 genes at full-length. Interestingly, when allowing the CEGMA proteins to be split over two scaffolds, the ratio raised to 91% (226 out of 248 genes). In comparison, most of the published lepidopteran genomes available in Lepbase contain more than 94% of the corresponding conserved CEGMA genes at least partially. This proportion is lower in four species, namely ca. 93% for *Heliconius erato demophoon*, 92% for *Plutella xylostella*, 84% for *Melitaea cinxia* and 62% for *Chilo suppressalis* (Table 2). As for the BUSCO analyses, 34% and 47% of the conserved eukaryote and arthropod, respectively, were identified in the *Tpit*-SP v1 genome assembly. These proportions rose to 57% (eukaryotes) and 72% (arthropods) when genes were allowed to be split on two scaffolds. In comparison, in the high quality genome assemblies of *S. frugiperda* strains 87% (corn) and 92% (rice) of the BUSCO arthropod gene set were identified (Gouin et al., 2017).

All genome characteristics obtained in the present study indicated that the *Tpit*-SP v1 genome assembly is of acceptable completeness and that the contigs are of good quality. Yet, the scaffolding is not optimized and the assembly is still fragmented. However, the generated data permitted to develop a range of population genomics approaches such as RADseq or SNP identification from genome-wide resequencing (e.g., Leblois et al., 2018, Gautier et al.,

pers. comm.), and therefore represents a valuable resource. The high fragmentation still does not allow analyses of linkage disequilibrium, haplotypic data and comparative analyses over large scale (e.g. synteny). Hence, the genome still needs to be improved in the future, with a particular focus on scaffolding and thus the generation and the integration of long reads.

Construction and quality assessment of *Tpit*-SP transcriptomic resources

Building a predicted gene set from the Tpit-*SP v1 genome*

In the genome assembly 29 415 predicted coding genes (Aug2.1) were identified which is close to the 26 329 predicted genes for the rice variant of *S. frugiperda* (Table 2). Within the available lepidopteran nuclear genomes, the average number of predicted coding genes is 17 157, with a minimum of 10 117 and a maximum of 26 329 genes (Table 2). In Aug2.1, 30 860 transcripts (CDS) were predicted. Furthermore, 89 225 exons were identified with an average length of 176 bp. On average, a gene was composed of 3 exons, with a maximum of 22 exons. The CDS of the predicted coding genes were on average 511 bp long, and summed up to 15.8 Mb. The cumulated length of the Aug2.1 introns was 220.8 Mb and the cumulated length of the intergenic regions was 301 Mb. The predicted Aug2.2 subset of high quality coding genes consisted of 8 232 CDS which were retained to build the reference transcriptome.

De novo *assembly of HiSeq and MiSeq transcriptomes*

Regarding the *de novo* transcriptomes, the best assemblies (highest percentage of identified CEGMA genes) were obtained with k-mer values of 61 and 51 for the HiSeq and MiSeq transcriptomes, respectively. These two transcriptomes were thus kept and named HiSeq61 and MiSeq51, respectively. Around 86% (HiSeq61) and 70% (MiSeq51) of the set of

cleaned, merged and extended (FLASH) RNAseq reads were used to establish the final assemblies. The average read per base coverage was 72 for HiSeq61 and 37 for MiSeq51. The HiSeq61 transcriptome covered 128 Mb (including alternative transcripts) and held 62 376 sequences which were grouped into 31 648 unigenes (Table 5). The MiSeq51 assembly had a total size of 152 Mb and consisted of 63 175 transcripts regrouped into 22 412 unigenes. The N50 values of the assembled HiSeq61 and MiSeq51 transcripts were 4177 and 3930 bp, respectively.

Reference transcriptome from the de novo *transcriptomes and predicted gene set: construction, quality assessment and localization*

The identification of coding sequences in transcripts with FrameDP resulted in 31 415 HiSeq61, 36 547 MiSeq51 *de novo* transcripts and in 6486 Aug2.2 genes with predicted CDS. This set was complemented by a previously published 454/Sanger SP transcriptome (Gschloessl et al., 2014) which had 5830 transcripts with predicted CDS, summing up to a total of 80 278 transcripts with CDS. Then, the longest FrameDP-peptide to each transcript was recovered and all protein sequences were subsequently clustered with CD-HIT into a reference protein set of 29 701 peptides. Of these reference sequences 20 465 (69%) had a full-length predicted CDS. Most reference proteins, *i.e.* 19 518 (66%) sequences, were only present in one set, *i.e.* specifically reconstructed by a sequencing or prediction methodology, while 261 (1%) of these reference peptides were present in all four input protein sets, 4172 (14%) in three sets and 5750 (19%) in two sets (Fig. 3). Surprisingly, the HiSeq and MiSeq transcript assemblies showed a relatively low overlap of the corresponding protein sets. Precisely, 6606 HiSeq and 9540 MiSeq protein clusters contained sequences that were obtained by one of these technologies only, while 8754 clusters were shared among HiSeq and MiSeq results. While we did not clearly identified the cause of this low overlap, we

suggest that sequencing specificities of HiSeq and Miseq technologies might explain these differences. These results emphasize the necessary cautions in future studies before combining data issued from different sequencing technologies. Yet, the filtering procedure we applied to various data sets prior to the construction of the reference transcriptome was meant to lower potential biases and to retain only highly reliable transcripts and genes. This is underlined by the high quality of the final reference transcriptome, which identified 246 out of 248 CEGMA eukaryotic core genes as being at least partially present and 238 at full length (Table 5). Concerning the 429 BUSCO conserved eukaryotic genes, 378 (88%) genes were found of which 350 were present at full length. The number of identified BUSCO arthropod genes within the reference transcriptome was 2236 out of 2675 (84%), 1938 being recovered at full-length. 24 648 (83%) reference transcripts matched with the Lepbase reference protein set. Of these, 13 802 reference transcripts had OHR values of at least 0.6, while 9578 reference transcripts had OHRs of at least 0.9, hence covering the respective best Lepbase protein sequence hits at least at 90% of their length. Overall, as can be seen in Table 5, compared to the three *de novo* transcriptomes and the Aug2.2 CDS set, the CD-HIT clustering approach generated a *Tpit*-SP reference transcriptome of higher quality than any of the four sets taken separately. Hence, by combining various approaches and sequencing technologies, the *Tpit*-SP transcript set was significantly optimized. When the 29 701 reference transcripts were aligned against the *Tpit*-SP v1 genome assembly, 9604 (32%) transcripts were located with 92% sequence identity on one single genome scaffold. Among the other transcripts, 17 149 (58%) were identified as being split on two scaffolds. In total, 26 753 (90%) reference transcripts could be located in the genome assembly. The structural and functional gene annotation of the genomic resource was greatly improved by positioning these actual (not predicted) transcripts. Hence, the reference transcriptome, along with the genome draft, will constitute an important resource to interpret e.g. outlier SNPs in the

population genomic studies, by providing valuable and annotated information in their vicinity.

Functional annotation and orthology analyses

Among the 29 701 reference transcripts, 71% (n=20 989) could be functionally annotated. This proportion is higher than those reported for *S. frugiperda* (42%), *Spodoptera litura* (38%) and *Spodoptera exigua* (40%) (Legeai et al., 2014; Pascual et al., 2012; Song et al., 2016). Of the annotated reference transcripts 13 907 (47%) were associated to 3896 distinct GO terms. The category 'biological process' comprised 2053 GO terms (15 205 assignments, Table S4), whereas 1275 GO terms belonged to 'molecular function' (18 361 assignments, Table S5) and 568 to 'cellular component' (10 609 assignments, Table S6). The three most represented biological processes were oxidation-reduction process (n=723), proteolysis (n=502) and regulation of transcription (n=500). ATP binding (n=1385), metal ion binding (n=873) and zinc ion binding (n=827) were the most frequent molecular function assignments. Concerning the GO category cellular component, integral component of membrane (n=3355), nucleus (n=913) and membrane (n=776) were the most represented terms.

The orthology study aimed to identify counterparts of *T. pityocampa* reference transcripts in the lepidopter species *S. frugiperda*, *B. mori* and *D. plexippus* as well as in the fruit fly *D. melanogaster*. Hence, this analysis contributed to the biological validation of the reference coding gene set and represented an essential indicator for its quality. In total, OrthoMCL identified 16 779 groups for the five species taken together (Fig. 4). 16 743 (56%) among the PPM reference transcripts with predicted proteins were assigned to 11 187 ortholog groups. This number of unique genes is lower than the 19 471 ortholog groups reported for the *S.*

frugiperda genome strains (supplementary data in Gouin et al., 2017). A majority of the PPM orthologs (93%, n=10 355) were shared with at least one of the other lepidopteran species. 413 PPM orthologs (528 transcripts) were specific to both Noctuoidea species (*S. frugiperda* and *T. pityocampa*) and 774 orthologs – corresponding to 2675 protein-coding transcripts - were highly divergent from the other four insect species.

CONCLUSIONS

This study provides the first genome assembly of a Notodontidae species, and the second for the species-rich Noctuoidea super-family which is currently underrepresented in the public databases. Right now (October 2017), the majority (n=13) of the 21 lepidopteran genomes publically available in Lepbase belong to the Papilionoidea super-family.

In addition to the *Tpit* draft genome, a high-quality reference transcriptome for this species is released. Thus, the resources developed in this study for the PPM will be of prime importance for Lepidoptera comparative genomics and transcriptomics which will shed light on this neglected taxonomic clade. In the case of the PPM and in particular the evolution of the phenologically shifted SP in Portugal, the availability of this draft genome will allow genome-wide analyses of genetic diversity to disentangle its recent evolutionary history (Leblois et al., 2018), and to gain knowledge on the genetic architecture of major adaptive traits such as phenology or tolerance to high temperature, even though the so far limited scaffolding still impedes some applications requiring additional genomic information. The *de novo* transcriptomic resources will permit a thorough analysis of major gene families involved in adaptation to climatic conditions (e.g., heat shock proteins) or biotic interactions (e.g., P450 involved in host use, immune reactions linked to interactions with natural enemies). They will also allow comparative transcriptomics and differential expression

studies using RNAseq approaches across developmental stages or populations and contribute to annotate functionally identified SNPs.

As shown by several quality measurements, the obtained contigs were characterized by good coverage and completeness values, while the scaffolding is still to be improved. One of the main limitations for building a genome-wide assembly in PPM is that inbred lines were not available due to heavy mortality in rearing conditions (Branco M.R., pers. comm.) and the very urticating L5 larval stage. The samples considered here, although originating from a population with a depleted genetic diversity, still remained sub-optimal for genome assembly because of its intrinsic heterozygosity. Yet, molecular and algorithmic alternatives can be developed to improve future genome assemblies. For example, adding long-fragment sequences, such as those produced by PacBio or MinIon technologies (Flusberg et al., 2010; Giordano et al., 2017) in the assembly process has been proven to facilitate the scaffolding process and hence raise the overall quality of a *de novo* genome (Rhoads & Au, 2015). In addition, as shown in the current *Tpit*-SP v1 assembly, around 58% (n=17 149) of the established reference transcripts were split on two separate genome scaffolds and 2 948 (10%) transcripts could not be localized at all on the genome. Recent studies (M. Chen et al., 2015; Xue et al., 2013) proposed to embed transcriptomic data into the assembly to improve recovery of coding regions and hence to increase the quality of scaffolds and coding regions. Last, building a linkage map could further contribute to associate and order BACs, genes and other markers to specific super-scaffolds. A future, improved *T. pityocampa* reference genome would allow to establish and refine studies of genetic diversity either to identify signature of selections through genomic scans (Gautier, 2015; Vitalis, Gautier, Dawson, & Beaumont, 2014) or to develop powerful analyses of demographic scenarios using a large amount of neutral SNPs and linkage disequilibrium data.

ACKNOWLEDGEMENTS

This work was supported by a grant from the French National Agency for Research (ANR-10-JCJC-1705-01 - GENOPHENO) and by the Institut National de la Recherche Agronomique (GAPP project, INRA AIP BioRessources 2012). Work in Portugal was partially supported by the project UID/AGR/00239/2013.

Karyotyping experiments were performed at the "Génomique Environnemental/Cytogénomique Evolutive", and DNA and RNA sonication was performed at the GenSeq technical facilities, both from the LabEx CeMEB (Laboratoire d'Excellence Centre Méditerranéen de l'Environnement et de la Biodiversité, Montpellier, France). The authors acknowledge the staff from the GPTR genotyping platform of Cirad Montpellier (France) for their support in the development of "genotyping applications" (MiSeq sequencing). Most of the bioinformatics analyses were performed on the CBGP HPC computational platform (Montferrier-sur-Lez, France). The authors would also like to thank the bioinformatics platforms ABiMS (Roscoff, France) and GenOuest (Rennes, France) for providing access to their computational resources and Fabrice Legeai (IGEPP-BIPAA, INRA, INRIA/Irisa, Campus Beaulieu, Rennes, France) for his bioinformatics advices.

References

Ahola, V., Lehtonen, R., Somervuo, P., Salmela, L., Koskinen, P., Rastas, P., . . . Hanski, I. (2014). The Glanville fritillary genome retains an ancient karyotype and reveals selective chromosomal fusions in Lepidoptera. *Nature Communications, 5*(4737). doi: 10.1038/ncomms5737

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., & Lipman, D. J. (1990). Basic local alignment search tool. *Journal of Molecular Biology, 215*(3), 403-410. doi: 10.1016/s0022-2836(05)80360-2

Bailly-Bechet, M., Haudry, A., & Lerat, E. (2014). "One code to find them all": A perl tool to conveniently parse RepeatMasker output files. *Mobile DNA, 5*(13). doi: 10.1186/1759-8753-5-13

Barnett, D. W., Garrison, E. K., Quinlan, A. R., Stromberg, M. P., & Marth, G. T. (2011). BamTools: A C++ API and toolkit for analyzing and managing BAM files. *Bioinformatics, 27*(12), 1691-1692. doi: 10.1093/bioinformatics/btr174

Battisti, A., Avcı, M., Avtzis, D. N., Ben Jamaa, M. L., Berardi, L., Berretima, W., . . . Zamoum, M. (2015). Natural history of the processionary moths (*Thaumetopoea spp.*): new insights in relation to climate change. In A. Roques (Ed.), *Processionary moths and climate change: an update* (pp. 15-80): Springer / Quae Editions.

Battisti, A., Holm, G., Fagrell, B., & Larsson, S. (2011). Urticating hairs in arthropods: their nature and medical significance. *Annual Review of Entomology, 56*, 203-220. doi: 10.1146/annurev-ento-120709-144844

Battisti, A., Larsson, S., & Roques, A. (2017). Processionary moths and associated urtication risk: global change-driven effects. *Annual Review of Entomology, 62*, 323-342. doi: 10.1146/annurev-ento-031616-034918

Battisti, A., Stastny, M., Netherer, S., Robinet, C., Schopf, A., Roques, A., & Larsson, S. (2005). Expansion of geographic range in the pine processionary moth caused by increased winter temperatures. *Ecological Applications, 15*(6), 2084-2096. doi: 10.1890/04-1903

Berardi, L., Pivato, M., Arrigoni, G., Mitali, E., Trentin, A. R., Olivieri, M., . . . Masi, A. (2017). Proteome analysis of urticating setae from *Thaumetopoea pityocampa* (Lepidoptera: Notodontidae). *Journal of Medical Entomology,* 54(6), 1560-1566. doi: doi:10.1093/jme/tjx144

Blankenberg, D., Von Kuster, G., Coraor, N., Ananda, G., Lazarus, R., Mangan, M., . . . Taylor, J. (2010). Galaxy: A web-based genome analysis tool for experimentalists. *Current Protocols in Molecular Biology,* 19, 1021. doi: 10.1002/0471142727.mb1910s89

Boetzer, M., Henkel, C. V., Jansen, H. J., Butler, D., & Pirovano, W. (2011). Scaffolding pre-assembled contigs using SSPACE. *Bioinformatics, 27*(4), 578-579. doi: 10.1093/bioinformatics/btq683

Bolger, A. M., Lohse, M., & Usadel, B. (2014). Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics, 30*(15), 2114-2120. doi: 10.1093/bioinformatics/btu170

Branco, M., Paiva, M. R., Santos, H. M., Burban, C., & Kerdelhué, C. (2017). Experimental evidence for heritable reproductive time in 2 allochronic populations of pine processionary moth. *Insect Science, 24*(2), 325-335. doi: 10.1111/1744-7917.12287

Burban, C., Gautier, M., Leblois, R., Landes, J., Santos, H., Paiva, M.-R., . . . Kerdelhué, C. (2016). Evidence for low-level hybridization between two allochronic populations of the pine processionary moth, *Thaumetopoea pityocampa* (Lepidoptera: Notodontidae). *Biological Journal of the Linnean Society, 119*(2), 311-328. doi: 10.1111/bij.12829

Bushnel, B. (2014). BBMap short read aligner, and other bioinformatic tools (Version 35.21). Retrieved from https://sourceforge.net/projects/bbmap/

Chalhoub, B., Belcram, H., & Caboche, M. (2004). Efficient cloning of plant genomes into bacterial artificial chromosome (BAC) libraries with larger and more uniform insert size. *Plant Biotechnology Journal, 2*(3), 181-188. doi: 10.1111/j.1467-7652.2004.00065.x

Challis, R. J., Kumar, S., Dasmahapatra, K. K. K., Jiggins, C. D., & Blaxter, M. (2016). Lepbase: The Lepidopteran genome database. *bioRxiv*. doi: 10.1101/056994

Chen, H., & Boutros, P. C. (2011). VennDiagram: A package for the generation of highly-customizable Venn and Euler diagrams in R. *BMC Bioinformatics,* 12, 35. doi: 10.1186/1471-2105-12-35

Chen, M., Hu, Y., Liu, J., Wu, Q., Zhang, C., Yu, J., . . . Wu, J. (2015). Improvement of genome assembly completeness and identification of novel full-length protein-coding genes by RNA-seq in the giant panda genome. *Scientific Reports, 5*(18019). doi: 10.1038/srep18019

Conesa, A., Götz, S., Garcia-Gomez, J. M., Terol, J., Talon, M., & Robles, M. (2005). Blast2GO: A universal tool for annotation, visualization and analysis in functional genomics research. *Bioinformatics, 21*(18), 3674-3676. doi: 10.1093/bioinformatics/bti610

Cong, Q., Borek, D., Otwinowski, Z., & Grishin, N. V. (2015). Tiger swallowtail genome reveals mechanisms for speciation and caterpillar chemical defense. *Cell Reports, 10*(6), 910-919. doi: 10.1016/j.celrep.2015.01.026

Derks, M. F., Smit, S., Salis, L., Schijlen, E., Bossers, A., Mateman, C., . . . Megens, H. J. (2015). The genome of winter moth (*Operophtera brumata*) provides a genomic perspective on sexual dimorphism and phenology. *Genome Biology and Evolution, 7*(8), 2321-2332. doi: 10.1093/gbe/evv145

Du, L., Li, W., Fan, Z., Shen, F., Yang, M., Wang, Z., . . . Zhang, X. (2015). First insights into the giant panda (*Ailuropoda melanoleuca*) blood transcriptome: A resource for novel gene loci and immunogenetics. *Molecular Ecology Resources, 15*(4), 1001-1013. doi: 10.1111/1755-0998.12367

El Mokhefi, M., Kerdelhué, C., Burban, C., Battisti, A., Chakali, G., & Simonato, M. (2016). Genetic differentiation of the pine processionary moth at the southern edge of its range: Contrasting patterns between mitochondrial and nuclear markers. *Ecology and Evolution, 6*(13), 4274-4288. doi: 10.1002/ece3.2194

Fitak, R. R., Mohandesan, E., Corander, J., & Burger, P. A. (2016). The *de novo* genome assembly and annotation of a female domestic dromedary of North African origin. *Molecular Ecology Resources, 16*(1), 314-324. doi: 10.1111/1755-0998.12443

Flusberg, B. A., Webster, D. R., Lee, J. H., Travers, K. J., Olivares, E. C., Clark, T. A., . . . Turner, S. W. (2010). Direct detection of DNA methylation during single-molecule, real-time sequencing. *Nature Methods, 7*(6), 461-465. doi: 10.1038/nmeth.1459

Fu, L., Niu, B., Zhu, Z., Wu, S., & Li, W. (2012). CD-HIT: Accelerated for clustering the next-generation sequencing data. *Bioinformatics, 28*(23), 3150-3152. doi: 10.1093/bioinformatics/bts565

Gautier, M. (2015). Genome-wide scan for adaptive divergence and association with population-specific covariates. *Genetics, 201*(4), 1555-1579. doi: 10.1534/genetics.115.181453

Giordano, F., Aigrain, L., Quail, M. A., Coupland, P., Bonfield, J. K., Davies, R. M., . . . Jackson, D. K. (2017). *De novo* yeast genome assemblies from MinION, PacBio and MiSeq platforms. *Scientific Reports, 7*(1), 3935. doi: 10.1038/s41598-017-03996-z

Gonthier, L., Bellec, A., Blassiau, C., Prat, E., Helmstetter, N., Rambaud, C., . . . Quillet, M. C. (2010). Construction and characterization of two BAC libraries representing a deep-coverage of the genome of chicory (*Cichorium intybus L.*, Asteraceae). *BMC Research Notes*, 3, 225. doi: 10.1186/1756-0500-3-225

Gouin, A., Bretaudeau, A., Nam, K., Gimenez, S., Aury, J. M., Duvic, B., . . . Fournier, P. (2017). Two genomes of highly polyphagous lepidopteran pests (*Spodoptera frugiperda*, Noctuidae) with different host-plant ranges. *Scientific Reports, 7*(1), 11816. doi: 10.1038/s41598-017-10461-4

Gouzy, J., Carrère, S., & Schiex, T. (2009). FrameDP: Sensitive peptide detection on noisy matured sequences. *Bioinformatics, 25*(5), 670-671. doi: 10.1093/bioinformatics/btp024

Gregory, T. R., Nicol, J. A., Tamm, H., Kullman, B., Kullman, K., Leitch, I. J., . . . Bennett, M. D. (2007). Eukaryotic genome size databases. *Nucleic Acids Research, 35*(Database issue), D332-D338. doi: 10.1093/nar/gkl828

Gschloessl, B., Vogel, H., Burban, C., Heckel, D., Streiff, R., & Kerdelhué, C. (2014). Comparative analysis of two phenologically divergent populations of the pine processionary moth (*Thaumetopoea pityocampa*) by *de novo* transcriptome sequencing. *Insect Biochemistry and Molecular Biology, 46*, 31-42. doi: 10.1016/j.ibmb.2014.01.005

Harris, R. S. (2007). *Improved pairwise alignment of genomic DNA.* (PhD thesis), The Pennsylvania State University, USA. Retrieved from http://www.bx.psu.edu/~rsharris/rsharris_phd_thesis_2007.pdf

Hodar, J. A., Zamora, R., & Castro, J. (2002). Host utilisation by moth and larval survival of pine processionary caterpillar *Thaumetopoea pityocampa* in relation to food quality in three *Pinus* species. *Ecological Entomology, 27*, 292-301. doi: 10.1046/j.1365-2311.2002.00415.x

International Silkworm Genome Consortium (2008). The genome of a lepidopteran model insect, the silkworm *Bombyx mori*. *Insect Biochemistry and Molecular Biology, 38*(12), 1036-1045. doi: 10.1016/j.ibmb.2008.11.004

Jakubowska, A. K., Nalcacioglu, R., Millan-Leiva, A., Sanz-Carbonell, A., Muratoglu, H., Herrero, S., & Demirbag, Z. (2015). In search of pathogens: Transcriptome-based identification of viral sequences from the pine processionary moth (*Thaumetopoea pityocampa*). *Viruses, 7*(2), 456-479. doi: 10.3390/v7020456

Kent, W. J. (2002). BLAT--the BLAST-like alignment tool. *Genome Research, 12*(4), 656-664. doi: 10.1101/gr.229202

Kerdelhué, C., Zane, L., Simonato, M., Salvato, P., Rousselet, J., Roques, A., & Battisti, A. (2009). Quaternary history and contemporary patterns in a currently expanding species. *BMC Evolutionary Biology,* 9, 220. doi: 10.1186/1471-2148-9-220

Kim, D., Pertea, G., Trapnell, C., Pimentel, H., Kelley, R., & Salzberg, S. L. (2013). TopHat2: Accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biology, 14*(R36). doi: 10.1186/gb-2013-14-4-r36

Krzywinski, M., Schein, J., Birol, I., Connors, J., Gascoyne, R., Horsman, D., . . . Marra, M. A. (2009). Circos: An information aesthetic for comparative genomics. *Genome Research, 19*(9), 1639-1645. doi: 10.1101/gr.092759.109

Langmead, B., & Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nature Methods, 9*(4), 357-359. doi: 10.1038/nmeth.1923

Leblois, R., Gautier, M., Rohfritsch, A., Foucaud, J., Burban, C., Galan, M., . . . Kerdelhué, C. (2018). Deciphering the demographic history of allochronic differentiation in the pine processionary moth *Thaumetopoea pityocampa*. *Molecular Ecology, accepted*.

Legeai, F., Gimenez, S., Duvic, B., Escoubas, J. M., Gosselin Grenet, A. S., Blanc, F., . . . Fournier, P. (2014). Establishment and analysis of a reference transcriptome for *Spodoptera frugiperda*. *BMC Genomics,* 15, 704. doi: 10.1186/1471-2164-15-704

Li, B., & Dewey, C. N. (2011). RSEM: Accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics,* 12, 323. doi: 10.1186/1471-2105-12-323

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., . . . 1000 Genome Project Data Processing Subgroup (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics, 25*(16), 2078-2079. doi: 10.1093/bioinformatics/btp352

Li, L., Stoeckert, C. J., & Roos, D. S. (2003). OrthoMCL: Identification of ortholog groups for eukaryotic genomes. *Genome Research, 13*(9), 2178-2189. doi: 10.1101/gr.1224503

Lukhtanov, V. A. (2014). Chromosome number evolution in skippers (Lepidoptera, Hesperiidae). *Comparative Cytogenetics, 8*(4), 275-291. doi: 10.3897/CompCytogen.v8i4.8789

Magoc, T., & Salzberg, S. L. (2011). FLASH: Fast length adjustment of short reads to improve genome assemblies. *Bioinformatics, 27*(21), 2957-2963. doi: 10.1093/bioinformatics/btr507

Manel, S., Perrier, C., Pratlong, M., Abi-Rached, L., Paganini, J., Pontarotti, P., & Aurelle, D. (2016). Genomic resources and their influence on the detection of the signal of positive selection in genome scans. *Molecular Ecology, 25*(1), 170-184. doi: 10.1111/mec.13468

Marcais, G., & Kingsford, C. (2011). A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics, 27*(6), 764-770. doi: 10.1093/bioinformatics/btr011

Mueller, J. C., Kuhl, H., Timmermann, B., & Kempenaers, B. (2016). Characterization of the genome and transcriptome of the blue tit *Cyanistes caeruleus*: Polymorphisms, sex-biased expression and selection signals. *Molecular Ecology Resources, 16*(2), 549-561. doi: 10.1111/1755-0998.12450

Mungall, C. J., & Emmert, D. B. (2007). A Chado case study: An ontology-based modular schema for representing genome-associated biological information. *Bioinformatics, 23*(13), i337-i346. doi: 10.1093/bioinformatics/btm189

NCBI Resource Coordinators. (2016). Database resources of the National Center for Biotechnology Information. *Nucleic Acids Research, 44*(Database issue), D7-D19. doi: 10.1093/nar/gkv1290

O'Neil, S. T., Dzurisin, J. D., Carmichael, R. D., Lobo, N. F., Emrich, S. J., & Hellmann, J. J. (2010). Population-level transcriptome sequencing of nonmodel organisms *Erynnis propertius* and *Papilio zelicaon*. *BMC Genomics,* 11, 310. doi: 10.1186/1471-2164-11-310

Parra, G., Bradnam, K., & Korf, I. (2007). CEGMA: A pipeline to accurately annotate core genes in eukaryotic genomes. *Bioinformatics, 23*(9), 1061-1067. doi: 10.1093/bioinformatics/btm071

Pascual, L., Jakubowska, A. K., Blanca, J. M., Canizares, J., Ferre, J., Gloeckner, G., . . . Herrero, S. (2012). The transcriptome of *Spodoptera exigua* larvae exposed to different types of microbes. *Insect Biochemistry and Molecular Biology, 42*(8), 557-570. doi: 10.1016/j.ibmb.2012.04.003

Peterson, D. G., Tomkins, J. P., Frisch, D. A., Wing, R. A., & Paterson, A. H. (2000). Construction of plant bacterial artificial chromosome (BAC) libraries: An illustrated guide. *Journal of Agricultural Genomics, 5*, 1-3.

Pimentel, C., Calvão, T., Santos, M., Ferreira, C., Neves, M., & Nilsson, J. Å. (2006). Establishment and expansion of a *Thaumetopoea pityocampa* (Den. & Schiff.) (Lep. Notodontidae) population with a shifted life cycle in a production pine forest, Central-Coastal Portugal. *Forest Ecology and Management, 233*(1), 108-115. doi: 10.1016/j.foreco.2006.06.005

Popescu, P., Hayes, H., & Dutrillaux, B. (1998). Techniques de cytogénétique animale. *Techniques et Pratiques* (pp. 260). Paris, FRA: INRA Editions.

Quinlan, A. R., & Hall, I. M. (2010). BEDTools: A flexible suite of utilities for comparing genomic features. *Bioinformatics, 26*(6), 841-842. doi: 10.1093/bioinformatics/btq033

Rhoads, A., & Au, K. F. (2015). PacBio sequencing and its applications. *Genomics Proteomics Bioinformatics, 13*(5), 278-289. doi: 10.1016/j.gpb.2015.08.002

Robinet, C., Imbert, C.-E., Rousselet, J., Sauvard, D., Garcia, J., Goussard, F., & Roques, A. (2012). Human-mediated long-distance jumps of the pine processionary moth in Europe. *Biological Invasions, 14*(8), 1557-1569. doi: 10.1007/s10530-011-9979-9

Robinet, C., Laparie, M., & Rousselet, J. (2015). Looking beyond the large scale effects of global change: Local phenologies can result in critical heterogeneity in the pine processionary moth. *Frontiers in Physiology, 6*(334). doi: 10.3389/fphys.2015.00334

Robinet, C., Rousselet, J., & Roques, A. (2014). Potential spread of the pine processionary moth in France: Preliminary results from a simulation model and future challenges. *Annals of Forest Science, 71*(2), 149-160. doi: 10.1007/s13595-013-0287-7

Rossi, J.-P., Garcia, J., Roques, A., & Rousselet, J. (2016). Trees outside forests in agricultural landscapes: Spatial distribution and impact on habitat connectivity for forest organisms. *Landscape Ecology, 31*(2), 243-254. doi: 10.1007/s10980-015-0239-8

Rousselet, J., Zhao, R., Argal, D., Simonato, M., Battisti, A., Roques, A., & Kerdelhué, C. (2010). The role of topography in structuring the demographic history of the pine processionary moth, *Thaumetopoea pityocampa* (Lepidoptera: Notodontidae). *Journal of Biogeography, 37*(8), 1478-1490. doi: 10.1111/j.1365-2699.2010.02289.x

Santos, H., Burban, C., Rousselet, J., Rossi, J. P., Branco, M., & Kerdelhué, C. (2011). Incipient allochronic speciation in the pine processionary moth *Thaumetopoea pityocampa* (Lepidoptera, Notodontidae). *Journal of Evolutionary Biology, 24*(1), 146-158. doi: j.1420-9101.2010.02147.x

Santos, H., Paiva, M. R., Tavares, C., Kerdelhué, C., & Branco, M. (2011). Temperature niche shift observed in a Lepidoptera population under allochronic divergence. *Journal of Evolutionary Biology, 24*(9), 1897-1905. doi: 10.1111/j.1420-9101.2011.02318.x

Schmieder, R., & Edwards, R. (2011). Quality control and preprocessing of metagenomic datasets. *Bioinformatics, 27*(6), 863-864. doi: 10.1093/bioinformatics/btr026

Schulz, M. H., Zerbino, D. R., Vingron, M., & Birney, E. (2012). Oases: Robust *de novo* RNA-seq assembly across the dynamic range of expression levels. *Bioinformatics, 28*(8), 1086-1092. doi: 10.1093/bioinformatics/bts094

Simao, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V., & Zdobnov, E. M. (2015). BUSCO: Assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics, 31*(19), 3210-3212. doi: 10.1093/bioinformatics/btv351

Skinner, M. E., Uzilov, A. V., Stein, L. D., Mungall, C. J., & Holmes, I. H. (2009). JBrowse: A next-generation genome browser. *Genome Research, 19*(9), 1630-1638. doi: 10.1101/gr.094607.109

Smit, A. F. A., & Hubley, R. (2008-2015). RepeatModeler (Version open-1.0). Retrieved from http://www.repeatmasker.org

Smit, A. F. A., Hubley, R., & Green, P. (2013-2015). RepeatMasker (Version open-4.0). Retrieved from http://www.repeatmasker.org

Song, F., Chen, C., Wu, S., Shao, E., Li, M., Guan, X., & Huang, Z. (2016). Transcriptional profiling analysis of *Spodoptera litura* larvae challenged with Vip3Aa toxin and possible involvement of trypsin in the toxin activation. *Scientific Reports, 6*(23861). doi: 10.1038/srep23861

Stanke, M. (2009). Incorporating RNA-Seq into AUGUSTUS. Retrieved 2017, from http://augustus.gobics.de/binaries/readme.rnaseq.html

Stanke, M. (2014). Incorporating Illumina RNAseq into AUGUSTUS with Tophat. Retrieved 2017, from http://bioinf.uni-greifswald.de/bioinf/wiki/pmwiki.php?n=IncorporatingRNAseq.Tophat

Stanke, M., & Morgenstern, B. (2005). AUGUSTUS: A web server for gene prediction in eukaryotes that allows user-defined constraints. *Nucleic Acids Research, 33*(Web Server issue), W465-W467. doi: 10.1093/nar/gki458

Stanke, M., & Waack, S. (2003). Gene prediction with a hidden Markov model and a new intron submodel. *Bioinformatics, 19*(Suppl. 2), ii215-ii225. doi: 10.1093/bioinformatics/btg1080

Suzek, B. E., Wang, Y., Huang, H., McGarvey, P. B., & Wu, C. H. (2015). UniRef clusters: A comprehensive and scalable alternative for improving sequence similarity searches. *Bioinformatics, 31*(6), 926-932. doi: 10.1093/bioinformatics/btu739

Tatusov, R. L., Fedorova, N. D., Jackson, J. D., Jacobs, A. R., Kiryutin, B., Koonin, E. V., . . . Natale, D. A. (2003). The COG database: An updated version includes eukaryotes. *BMC Bioinformatics,* 4, 41. doi: 10.1186/1471-2105-4-41

Vega, J. M., Moneo, I., Garcia-Ortiz, J. C., Gonzalez-Munoz, M., Ruiz, C., Rodriguez-Mahillo, A. I., . . . Vega, J. (2014). IgE sensitization to *Thaumetopoea pityocampa*: Diagnostic utility of a setae extract, clinical picture and associated risk factors. *International Archives of Allergy and Immunology, 165*(4), 283-290. doi: 10.1159/000369807

Vitalis, R., Gautier, M., Dawson, K. J., & Beaumont, M. A. (2014). Detecting and measuring selection from gene frequency data. *Genetics, 196*(3), 799-817. doi: 10.1534/genetics.113.152991

Vurture, G. W., Sedlazeck, F. J., Nattestad, M., Underwood, C. J., Fang, H., Gurtowski, J., & Schatz, M. C. (2017). GenomeScope: Fast reference-free genome profiling from short reads. *Bioinformatics*. doi: 10.1093/bioinformatics/btx153

Xue, W., Li, J. T., Zhu, Y. P., Hou, G. Y., Kong, X. F., Kuang, Y. Y., & Sun, X. W. (2013). L_RNA_scaffolder: Scaffolding genomes with transcripts. *BMC Genomics,* 14, 604. doi: 10.1186/1471-2164-14-604

Zdobnov, E. M., & Apweiler, R. (2001). InterProScan--an integration platform for the signature-recognition methods in InterPro. *Bioinformatics, 17*(9), 847-848. doi: 10.1093/bioinformatics/17.9.847

Zerbino, D. R., & Birney, E. (2008). Velvet: Algorithms for *de novo* short read assembly using de Bruijn graphs. *Genome Research, 18*(5), 821-829. doi: 10.1101/gr.074492.107

Zovi, D., Stastny, M., Battisti, A., & Larsson, S. (2008). Ecological costs on local adaptation of an insect herbivore imposed by host plants and enemies. *Ecology, 89*(5), 1388-1398. doi: 10.1890/07-0883.1

AUTHOR CONTRIBUTIONS

Conceived and designed the study: CK, MG, BG, RS. Performed sampling, insect rearing and ensured sample preservation: SR, MB. Performed wet lab experiments (karyotyping, DNA extraction, library construction, BAC construction, quality control, high-throughput sequencing): LS, CB, EL, OB, GB, HB, JN, SN, PG. Developed and performed bioinformatics analyses: BG, FD, AB, ED. Wrote the paper: BG, FD, CK. All authors read and approved the final manuscript.

DATA ACCESSIBILITY

The genome and transcriptome assemblies and all annotations are publicly available in the LepidoDB database (http://bipaa.genouest.org/sp/thaumetopoea_pityocampa/download/). The raw sequence reads as well as the draft genome, the BAC assemblies and the reference transcriptome are accessible as NCBI BioProject with the id PRJNA344465 (https://www.ncbi.nlm.nih.gov/bioproject/344465).

SUPPORTING INFORMATION

Additional Supporting Information may be found in the online version of this article:

Fig. S1 Metaphase obtained from an egg mass of *T. pityocampa*, allowing to determine 2n = 98 as the most likely number of chromosomes in that species.

Fig. S2 All 11 assembled BAC sequences and corresponding aligned *Tpit*-SP v1 genome scaffolds.

Table S1 Read counts for the libraries used for the *Tpit*-SP v1 genome assembly.

Table S2 Read per base coverage of paired-end reads on the *Tpit*-SP v1 genome scaffolds.

Table S3 Read counts for the libraries used for the BAC assemblies.

Table S4 GO term counts for the category 'biological process'.

Table S5 GO term counts for the category 'molecular function'.

Table S6 GO term counts for the category 'cellular component'.

FIGURE LEGENDS

Fig. 1 Workflow summarizing the bioinformatics analyses to generate the genomic and transcriptomic resources for *Tpit*-SP. A. Reconstruction of the 11 BAC sequences used for genome quality assessment; B. Assembly of the *Tpit*-SP v1 draft genome; C. Generation of the HiSeq, MiSeq and reference transcriptomes.

Fig. 2 Four chosen assembled BAC sequences and the corresponding aligned *Tpit*-SP v1 nuclear genome scaffolds.

Fig. 3 Venn diagram showing all reference transcripts and their coverage among the three transcriptome assemblies and the Aug2.2 CDS predictions.

Fig. 4 Venn diagram showing all OrthoMCL ortholog groups among the predicted *Tpit*-SP reference proteins (TPIT) and the proteomes of *S. frugiperda* (SFRU), *B. mori* (BMOR), *D. plexippus* (DPLE) and *D. melanogaster* (DMEL).

TABLES

Table 1 Features of the contigs and the scaffolds (≥ 1 kb) retained in the assembly of the nuclear *T. pityocampa* genome. The coverage is defined as the average read count per assembled bp.

	Contigs	Large scaffolds
Total length [Mb]	507	537
Sequence count	675 934	68 292
Mean [bp]	750	7870
Median [bp]	402	1757
N50 [bp]	1424	163 589
N50 sequence count	100 209	728
N90 [bp]	320	1951
N90 sequence count	379 839	29 439
Minimum length [bp]	51	1000
Maximum length [bp]	19 793	2 148 522
Second largest length [bp]	18 371	1 877 510
Total coverage	121	84
PE300i – coverage	76	52
PE600i – coverage	45	31
SE454 – coverage	0.9	0.5
GC content [%]	38.0	37.2
Count base N	0	116 266 296
N content in assembly [%]	0	21.6

Table 2 Characteristics of Lepidoptera genome assemblies available in Lepbase v4. Genome characteristics for *Spodoptera frugiperda* were taken from the corresponding publications Gouin et al. (2017) and Legeai et al. (pers. comm.). a) Counts refer to genes/transcripts predicted on the assembled genomes (Aug2.1 for *T. pityocampa*). b) The CEGMA % for *Melitaea cinxia* were missing in Lepbase and were therefore obtained from the corresponding genome paper (Ahola et al., 2014). c) In brackets the percentage of CEGMA genes is shown, including the genes which were split on two *Tpit*-SP v1 nuclear genome scaffolds.

Part A

Species	Super family; family	Common name	Version	Accession	Date	Size (Mb)	Scaffold count
Amyelois transitella	Pyraloidea; Pyralidae	Navel orangeworm	v1.0	GCA_001186105.1	2015-07-22	406	7301
Bicyclus anynana	Papilionoidea; Nymphalidae	Squinting bush brown	v1.2	N/A	2015-10-28	475	10 800
Bombyx mori	Bombycoidea; Bombycidae	Silkworm	v1.0	GCA_000151625.1	2008-04-28	482	43 463
Calycopis cecrops	Papilionoidea; Lycaenidae	Red-banded hairstreak	v1.1	GCA_001625245.1	2016-04-21	729	60 049
Chilo suppressalis	Pyraloidea; Crambidae	Rice striped stem borer	v1.0	GCA_000636095.1	2014-04-22	372	80 479
Danaus plexippus	Papilionoidea; Nymphalidae	Monarch butterfly	v3.0	N/A	2012-11-07	249	5397
Heliconius erato demophoon	Papilionoidea; Nymphalidae	Crimson-patched longwing	v1.0	N/A	2016-03-11	383	196
Heliconius erato lativitta	Papilionoidea; Nymphalidae	Crimson-patched longwing	v1.0	N/A	2016-09-12	418	142
Heliconius melpomene melpomene	Papilionoidea; Nymphalidae	Postman butterfly	v2.0	N/A	2015-07-17	275	795
Junonia coenia	Papilionoidea; Nymphalidae	Buckeye	v1.0	N/A	2017-05-13	586	1136
Lerema accius	Hesperioidea; Hesperiidae	Clouded skipper	v1.1	GCA_001278395.1	2015-09-02	298	29 988
Manduca sexta	Bombycoidea; Sphingidae	Tobacco hornworm	v1.0	GCA_000262585.1	2012-05-24	419	20 871
Melitaea cinxia	Papilionoidea; Nymphalidae	Glanville fritillary	v1.0	GCA_000716385.1	2015-04	390	8261
Operophtera brumata	Geometroidea; Geometridae	Winter moth	v1.0	GCA_001266575.1	2015-08-11	638	25 801
Papilio glaucus	Papilionoidea; Papilionidae	Eastern tiger	v1.1	GCA_000931545.1	2015-03-20	376	68 029

		swallowtail					
Papilio machaon	Papilionoidea; Papilionidae	Old world swallowtail	v1.0	GCA_001298355.1	2015-09-28	278	63 186
Papilio polytes	Papilionoidea; Papilionidae	Common Mormon	v1.0	GCA_000836215.1	2015-02-02	227	3873
Papilio xuthus	Papilionoidea; Papilionidae	Asian swallowtail	v1.0	GCA_001298345.1	2015-09-28	243	15 362
Phoebis sennae	Papilionoidea; Pieridae	Cloudless sulphur	v1.1	GCA_001586405.1	2016-03-10	345	20 800
Plodia interpunctella	Pyraloidea; Pyralidae	Indian meal moth	v1.0	N/A	2015-04-03	382	10 542
Plutella xylostella	Yponomeutoidea; Plutellidae	Diamondback moth	v1.1	GCA_000330985.1	2014-10-02	393	1794
Spodoptera frugiperda (corn strain)	Noctuoidea; Noctuidae	Fall armyworm	v3.1	PRJEB13110	2017-09-25	438	41 577
Spodoptera frugiperda (rice strain)	Noctuoidea; Noctuidae	Fall armyworm	v1.0	PRJEB13834	2017-09-25	371	29 127
Thaumetopoea pityocampa	Noctuoidea; Noctuidae	Pine processionnary moth	v1.0	PRJNA344465	2018-01	537	68 292

Part B

Species	N50 (kb)	N90 (kb)	CEGMA complete (%)	At least CEGMA partial (%)	GC%	N%	Gene count [a)]	Transcript count [a)]
A. transitella	1587.0	45.4	78.2	95.2	35.7	14.5	15 208	19 808
B. anynana	638.3	99.3	81.0	97.2	36.5	1.2	22 642	22 642
B. mori	4008.4	61.1	76.6	96.8	37.7	10.4	15 488	22 061
C. cecrops	233.5	4.8	70.6	94.8	37.1	5.5	16 456	16 456
C. suppressalis	5.2	2.4	41.5	61.7	35.7	12.5	10 117	10 132
D. plexippus	715.6	160.5	90.3	96.0	31.6	2.7	15 130	15 130
H. erato demophoon	10 689.0	2670.1	81.1	93.2	33.2	1.4	13 676	20 118
H. erato lativitta	5483.8	1432.2	*N/A*	95.0	33.5	1.3	14 613	14 613
H. melpomene melpomene	2102.7	273.1	88.7	96.8	32.8	0.4	20 102	21 661
J. coenia	1571.1	261.6	*N/A*	*N/A*	34.5	0.0	19 234	19 234
L. accius	525.3	60.3	83.9	95.2	34.4	2.9	17 411	17 411
M. sexta	664.0	46.4	85.9	96.0	35.3	4.7	15 451	27 403
M. cinxia	119.3	29.6	77.0 [b)]	83.9 [b)]	32.6	7.4	16 751	16 790
O. brumata	65.6	13.6	64.1	94.0	38.6	2.1	16 912	16 912
P. glaucus	230.3	2.0	84.3	96.0	35.4	3.6	15 692	15 692
P. machaon	1174.3	1.1	87.9	94.8	33.8	4.5	15 497	15 497
P. polytes	3672.3	930.4	83.9	94.0	34.0	3.9	12 244	12 244
P. xuthus	3432.6	22.4	91.5	96.4	34.1	5.4	15 322	15 322
P. sennae	256.7	19.6	82.3	96.0	33.0	3.2	16 117	16 492
P. interpunctella	1270.7	18.7	85.1	96.4	35.1	4.6	23 136	24 497
P. xylostella	737.2	154	78.2	92.3	38.3	14.4	19 386	23 907
S. frugiperda (corn strain)	52.8	3.5	*N/A*	*N/A*	36.0	2.6	21 700	21 779
S. frugiperda (rice	28.5	6.4	*N/A*	*N/A*	36.0	0.0	26 329	26 357

strain)								
T. pityocampa	163.6	2.0	35.1	23.4 (91.1) [c)]	37.2	21.6	29 415	30 860

Table 3 Number of repeated elements found in the *Tpit*-SP v1 genome and corresponding percentage of genome length.

Family	Fragments	Total length [Mb]	% of genome
LTR	702 636	105.3	19.6
LINE	364 879	58.9	11.0
SINE	331 115	50.4	9.4
DNA	174 994	28.7	5.3
Total	1 573 624	243.3	45.3

Table 4 Characteristics of the 11 sequenced and assembled BACs. The estimated and assembled sizes are given, as well as the k-mer length used for the assembly, the number of scaffolds, N50 lengths, BAC read coverages and the % of BAC sequence lengths covered by *Tpit*-SP v1 genome scaffolds, contigs and genomic reads.

PE BAC library	Estimated BAC size [kb]	Velvet + SSPACE [kb]	Best assembly (k-mer)	Scaffold count	N50 [kb]	Average coverage (reads/bp)	# Aligned *Tpit*-SP scaffolds	% covered by genome scaffolds	% covered by genome contigs	% covered by ≥ 10 genome reads/bp
Tpi21J02_S14_L001	40	49.4	181	2	26.4	1386	15	83.3	94.8	92.6
Tpi21G16_S15_L001	90	88.2	171	5	55.9	1249	24	41.4	67.4	88.2
Tpi21F03_S16_L001	85	95.8	171	1	95.8	1553	23	65.5	78.0	84.3
Tpi21A08_S17_L001	55	64.8	181	1	64.8	1398	15	73.5	84.4	89.7
Tpi21H19_S18_L001	70	102.6	181	2	98.1	1456	30	65.0	78.8	91.2
Tpi21L11_S19_L001	50	60.1	161	3	36.3	1484	21	52.2	74.6	86.7
Tpi21C18_S20_L001	75	83.4	171	2	63.8	1297	19	55.4	71.3	81.9
Tpi21D23_S21_L001	128	100.0	181	9	14.6	1202	11	14.1	33.6	71.1
Tpi21P07_S22_L001	65	46.8	181	2	42.5	2148	5	27.1	60.5	84.3
Tpi21K05_S23_L001	95	94.7	171	1	94.7	921	32	72.4	82.5	86.1
Tpi21M24_S24_L001	68	81.2	181	1	81.2	1541	22	68.9	83.2	92.0

Table 5 Characteristics of the various *Tpit*- SP transcriptome assemblies and the robust CDS subset (Aug2.2) of the coding-genes predicted from the genome. a) Reference transcriptome obtained by CD-HIT clustering: only unigenes are present (compared to presence of alternative transcripts in other transcriptomes) b) Including alternative transcripts. c) For Augustus Aug2.2 predictions the CDS lengths were considered. d) N50: contig length for which half of the assembly is represented by contigs of this size or longer.

	Reference [a)]	**Aug2.2**	**454/Sanger**	**SP HiSeq**	**SP MiSeq (MiSeq1+2)**
Raw number of NGS reads	*N/A*	*N/A*	467 082	559 510 744	44 533 734
Read number after cleaning	*N/A*	*N/A*	465 703	514 941 496	39 613 596
Read number after FLASH	*N/A*	*N/A*	*N/A*	457 553 121	30 669 751
Read number after BBNORM (only HiSeq)	*N/A*	*N/A*	*N/A*	119 171 986	*N/A*
Mean NGS trimmed read length [bp] (min–max)	*N/A*	*N/A*	308 (220–593)	91.1 (30-168)	273.4 (30-568)
Number of Sanger reads	*N/A*	*N/A*	5290	*N/A*	*N/A*
Mean Sanger trimmed read length [bp] (min–max)	*N/A*	*N/A*	515 (334–1832)	*N/A*	*N/A*
Size of transcriptome [Mb]	67.7	3.1	9.4 [b)]	128.3 [b)]	151.9 [b)]
Number of mapped cleaned (and FLASH+BBNORM) reads	*N/A*	*N/A*	309 200	102 164 686	21 300 953
Coverage (mean read count per bp)	*N/A*	*N/A*	17.1 [b)]	72.4 [b)]	37.3 [b)]
Number of unigenes	29 701	8232	6696	31 648	22 412
Number of transcripts	29 701	8232	8119 [b)]	62 376 [b)]	63 175 [b)]
Mean transcript length [bp]	2279	378 [c)]	1156	2057	2404
Median transcript length [bp]	1564	294 [c)]	1001	1130	1504
N50 transcript length [bp] [d)]	3632	420 [c)]	1386	4177	3930
Located on *Tpit*-SP v1 genome [count] (%)	9604 (32.3)	8232 (100)	3100 (38.2)	22 824 (36.6)	19 854 (31.4)
Split on two *Tpit*-SP v1 scaffolds [count] (%)	17 149 (57.7)	0 (0)	4196 (51.7)	31 845 (51.1)	36 476 (57.7)
CEGMA identified [%] (count of 248)	99.2 (246)	*N/A*	66.9 (166)	96.0 (238)	95.2 (236)

CEGMA full-length [%] (count of 248)	96.0 (238)	*N/A*	60.5 (150)	89.5 (222)	82.3 (204)
BUSCO euk identified [%] (count of 429)	88.1 (378)	*N/A*	52.4 (225)	86.9 (373)	86.0 (369)
BUSCO euk full-length [%] (count of 429)	81.6 (350)	*N/A*	43.1 (185)	80.0 (343)	72.3 (310)
BUSCO arthropod identified [%] (count of 2675)	83.6 (2236)	*N/A*	31.5 (843)	81.7 (2186)	80.0 (2141)
BUSCO arthropod full-length [%] (count of 2675)	72.4 (1938)	*N/A*	25.9 (692)	70.4 (1882)	60.5 (1618)
Transcripts with FrameDP peptide [count] (%)	29 701 (100)	6486 (78.8)	5830 (71.8)	31 415 (50.4)	36 547 (57.9)
Transcripts with FrameDP full-length peptides [count] (%)	20 465 (68.9)	2655 (32.3)	3987 (49.1)	24 651 (39.5)	27 314 (43.2)